MDS and clustering: an experimental pipeline to identify relevant characteristics in COVID-19 prognosis

Constanza Marini[1] and Diana Laura Aguilar[1]

Tecnologico de Monterrey, Carretera al Lago de Guadalupe Km. 3.5, Atizapán de Zaragoza, Estado de México 52926, México.
A01332485@itesm.mx
A01751168@itesm.mx

Abstract. COVID-19 is a recently identified acute respiratory syndrome that has taken the lives of many, puts in jeopardy global health, and challenges health systems all around the world. This outbreak has called upon the whole science community to help fight this disease. Motivated by this, in this paper, we present an experimental pipeline so as to extract information from a COVID-19 dataset. Initially, we leveraged multidimensional scaling to obtain lower-dimensional representations of our data and then performed cluster analysis. Later on, our analysis was divided into two fundamental lines as so was our original dataset. On the one hand, the analysis we conducted on a first subset yields that the probability of infection does not present important variations when analyzing cluster by cluster. On the other hand, we conducted one-way ANOVA tests on a second subset to find statistical variations among the clusters, and found the attributes that variate most.

Keywords: Clinical data · Clustering · COVID-19 · Probability of infection · SARS-CoV-2.

1 Introduction

Severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), also known as HCoV-19, is a newly recognized illness which has affected 6,640,641 people worldwide and has taken the lives of 389,965 as for June 4th, 2020 [1]. Until today, there are seven coronavirus which infect humans: SARS-CoV, MERS-CoV, SARS-CoV-2, HKU1, NL63, OC43 and 229E; the first three viruses can cause severe disease, while the latter are associated with mild symptoms. SARS-CoV-2 is optimized to bind the human receptor ACE2 with the receptor-binding domain within the spike protein. This domain has high affinity to ACE2 from humans, ferrets and cats which have receptor homology. The receptor-binding domain of SARS-CoV-2 is most likely the result of natural selection which allowed an optimal binding solution. Moreover, a subunit of the SARS-CoV-2 spike plays an important role in the viral infectivity and host range. Although the real source of

SARS-CoV-2 is unkown, the RaTG13 bat virus has 96% similarity to the coronavirus. On the other hand, malayan pangolins contain coronaviruses similar to SARS-CoV-2 receptor-binding domain [2].

Coronavirus disease (COVID-19) pneumonia ranges from mild to critically severe cases, characterized by severe hypoxaemia. In many regions during the outbreak of SARS-CoV-2, the number of patients who needed mechanical ventilation exceeded the capacity of the intensive care units [3]. According to the WHO global surveillance for COVID-19, a suspect case is any patient with acute respiratory illness presenting fever and at least one symptom of respiratory disease (cough, shortness of breath) [4]. However, a study performed in Wuhan Jin Yin-tan hospital, China, found that 11.5% of the patient cohort presented fever 2-8 days after the onset of the disease. The most common symptoms detected in the study were fever (98%), cough (77%), dypsnoea (63.5%) and all patients had bilateral infiltrates confirmed by chest X-rays. Moreover, patients developed organ function damage: acute respiratory distress syndrome (67%), acute kidney injury (29%), cardiac injury (23%) and liver dysfunction (29%). The median duration of patients in intensive care units from admission to death was 7 days. The non-survivors were men and people older than 65 years and were more likely to have a chronic medical illness. Also, those with COVID-19 developed lymphocytopenia; the severity of lymphocytopenia reflected the severity of the infection. SARS-CoV-2 demands a lot of critical care resources; if a hospital cannot poses the resources and highly prepared staff, the disease can posse a great threat [3].

The outbreak of SARS-CoV-2 represents an urgent threat to global health. However, the globalized society and sharing of scientific data offers a promising instrument to fight this illness. Only in four months, more than 12,400 articles about this virus have been published. Furthermore, machine learning (ML) has been employed to consider various hypothesis at a single experiment in several pathologies, and COVID-19 is not the exception. ML can alleviate the workload of medical experts by decreasing the time required to produce an analysis and allowing artificial intelligence practitioners to support clinicians [5]. Automated analysis of chest computed tomography scans can predict with high accuracy (96.78%) a patient with COVID-19 lesions [6]. Nonetheless, the power of these studies remains limited due to small cohorts and poorly controlled confounds. ML can also accelerate the screening of treatments by considering several potential antiviral agents and give a prediction of the drug-virus interaction based on DNA sequences and protein structure. Likewise, a large spectrum of vaccine candidates can be screened in the same way [5]. A very recent study leveraged a database of blood samples of patients in the region of Wuhan, China, to identify biomarkers of disease mortality. The authors used ML tools and selected three biomarkers that can predict the mortality of a COVID-19 patient more than 10 days before the event with more than 90% accuracy. This method allows the prediction of patients at the highest risk in order to prioritize the intensive care and reduce the mortality rate [7].

Although a plethora of research is being done on this subject, there is still much concern as to help fight this disease. In this framework, we present an experimental pipeline based on Multimensional Scaling (MDS) and clustering. We leveraged MDS to lower the dimension of our data, and then used clustering to extract intra- and inter-cluster information. As our original dataset was divided into two subsets, so was our study. On the one hand, we found that the probability of infection does not present high variations inside the clusters, but does change from cluster to cluster for Set 1. Moreover, we conducted one-way ANOVA tests to evaluate the statistical differences among clusters for Set 2. Our findings show that there exist attributes that vary from cluster to cluster.

The rest of this paper is organized as follows. Section 2 introduces two key concepts behind our research, i.e, MDS and clustering. Section 3 presents previous research related to ours briefly. Section 4 outlines our experimental pipeline, and shows our findings. Finally, Section 5 draws our conclusions.

2 Preliminars

Two main concepts set the basis for our research: clustering and dimensionality reduction. Thus, in this section, we describe the key ideas behind these topics.

2.1 Clustering

Unsupervised machine learning focuses on the identification of useful properties of available data without labeling. The most common task is to look up for groups of similar samples, named clustering. The centroids of these groups can be used as predictors of unknown attribute values, as visualization tools for multidimensional data and to generate higher-level attributes from the original data. Clustering analysis has as input a set of samples described by a vector of attributes but without class labels, as mentioned before; the output is a set of clusters of common samples. The clusters are described by their centroids, and the identification of a cluster must be given by a measuring distance from the sample to the centroid. Moreover, if the samples are described by numeric and discrete attributes, the distance can be obtained from the sum of squared distances along corresponding attributes [8].

There are several clustering algorithms which are based on different approaches. A simple algorithm is k-means, where k denotes the requested number of clusters defined by the users [8]; this approach has a good performance when the clusters are of similar size, density and have a globular shape [9]. On the other hand, hierarchical aggregation is based on the inter-cluster distance. First, each sample defines its own cluster, and as the process continues, the clusters are formed according to the smallest mutual distance [8]. Another approach is based on density. In this framework, a cluster is a set of points spread in the data space over a contiguous area of high density of data points. The density-based clusters separates from each other by low density regions; this low density regions are considered as noise or outliers. One of the main advantage of density-based clustering is that it can identify clusters with unusual shape ignoring noise [10].

2.2 Dimensionality reduction

Along with clustering, dimensionality reduction must be mentioned. Dimension reduction is the process of removing some attributes of a dataset by identifying a set of principal variables. Dimensionality reduction is desirable as it allows models to be more efficient in terms of execution time and it tends to increase the accuracy of the model [11] [12]. Principal component analysis (PCA) is the main linear method for this process. It generates a set of components that represents the most relevant information from the original data. The objective of PCA is to reduce the number of predictors and maximize the variance [9]. Another method is multidimensional scaling (MDS). MDS translates the information of the pairwise distances into an abstract Cartesian plane preserving this distances as intact as possible [13]. Moreover, MDS can follow similarity matrices based on different metrics since it employs the inter-element distances rather than the coordinates of the element. This method has proven to obtain good visualizations of global data associated with human pathologies [14].

3 Previous work

Similar approaches have been used to analyze COVID-19 data. Carrillo-Larco and Castillo-Cara in [9] developed an unsupervised model in order to cluster countries in groups with similar number of confirmed COVID-19 cases, based on pre-pandemic variables. The authors informed that the k-means method by disease prevalence estimates (diabetes, chronic obstructive pulmonary disease, tuberculosis and HIV/IDS), metrics of air pollution (particles of width<2.5 μm), socioeconomic status (probability of a person to adopt preventive care) and health system coverage (access to appropriate healthcare), which are country-level variables. These predictors were chosen since they have a close relation with COVID-19. Moreover, the predictors were reformed by an orthogonal transformation and with principal component analysis (PCA) specified for three components. The authors optimized a cost function in order to find the best number of clusters, i.e. the initialization of the centroids. For the statistical analysis, one-way ANOVA tests were done to compare the COVID-19 related variables between clusters; pairwise combinations within clusters were analyzed with adjusted t-tests with Bonferroni method. The conclusions in [9] were that the clusters had a strong difference regarding the order of the first confirmed case; the model may suggest that the number of cases in a country grouped in one cluster will be within the proposed range for that cluster.

Another study by Dolgikh in [15] analyzed the distribution of case data with the most informative predictors, selected by unsupervised machine learning methods, and identified the cases with the heaviest impact. The author employed PCA and dimensionality reduction with neural network autoencoder models in order to identify and separate classes that can be linked to the outcome. The observable parameters included genetic differences, population density, social traditions, immunization, smoking rate and epidemiological policy course. Moreover, Dolgikh generated two clusters by PCA: one identified as cases with relative high impact (above 0.8 mortality per capita per million of population), and the other defined as milder-impact cases. The cases of Italy, Spain and New York were grouped in the first cluster, while the cases of United Kingdom, France, Belgium and Netherlands were in the second cluster. The highest influence factors were policy time, connection hub, social proximity, immunization and the smoking habit. The autoencoder model indicated the same results as the PCA. The author concluded that the findings may be used in the evaluation of possible epidemiological scenarios and as a tool to identify the areas of potential risk.

Additionally, in [14], the authors proposed a methodology to extract mathematical information on infectious agents. Their methodology consists in using MDS to visualize the relationships among viral infectious diseases that affect humans. Initially, the observations were in a seven-dimensional space. Then, after applying MDS, they obtained representations in two and three dimensions. Later on, they proposed two clustering algorithms as a means to extract information from the MDS representation, that is, K-means and Hierarchical clustering. Their findings show that MDS can be adopted so as to represent information on viral diseases.

4 Experimental setup

In this section, we start by describing the data we utilized, the experimental pipeline we followed along with the results we obtained.

4.1 The data

The data used for this research was obtained from the public datasets platform of Kaggle; the name of the dataset is *Coronavirus disease 2019 (COVID-19) India* [16]. It includes clinical, personal and regional information for patients with probable COVID-19. This dataset comes with two sets, called from now on, Set 1 and Set 2. Set 1 contains 10,714 observations, 27 attributes, and the target is the infection probability of SARS-CoV-2. On the other hand, Set 2 contains 14,498 observations and the same attributes, but it does not include the target. In Table 1, we present a summary of the original attributes of the dataset.

In order to extract information from the data, we conducted a series of procedures to remove irrelevant features, impute missing values, and transform categorical attributes. First and foremost, we removed irrelevant features, i.e, the name and ID of the patients, and the designation. Then, as it is customary in health-care datasets [17], there was a plethora of missing values. Conducting a complete-case experiment could have resulted in omitting much information from the data, as suggested by [17]. Thus, we leveraged the *mice* package from R [18] to impute missing values. *mice* stands for Multivariate Imputation by Chained Equations, and it is one of the fundamental approaches for imputing multivariate data. What *mice* does is to impute each incomplete variable with a separate model.

Table 1. Attributes description of training/test sets: In the first column, we present the attributes; in the second, the variable type, either numerical or categorical, and finally, we present the attribute description.

Attribute	Variable type	Description
Patient ID	Num	Unique identification number for each patient
Region	Cat	Area that the patient belongs to
Gender	Cat	Gender of the patient
Designation	Cat	Designation of the patient
First_Name	Cat	Name of the patient
Married	Cat	Marital status of the patient
Children	Num	Number of children
Occupation	Cat	Sector of the patient's occupation
Mode_transport	Cat	Most used mode of transportation by the patient
Cases/1M	Num	Number of confirmed cases per million inhabitants in that region
Deaths/1M	Num	Number of deaths per million inhabitants in that region
Comorbidity	Cat	Presence of one or more additional medical conditions co-ocurring with a primary condition
Age	Num	Age of the patient
Coma score	Num	Neurological coma score
Pulmonary score	Cat	Pulmonary PaO_2/FiO_2
Cardiological pressure	Cat	Cardiological mean systolic arterial pressure (mmHg)
Diuresis	Num	Diuresis (mL/day)
Platelets	Num	Haematological platelets (10/L)
HBB	Num	Hepatic blood bilirubin (μmol/L)
D-dimer	Num	D-dimer concentration in the blood (ng/mL)
Heart rate	Num	Number of times the patient's heart beats per minute
HDL cholesterol	Num	High-density lipoprotein level (mg/dL)
Charlson Index	Num	One-year mortality for a patient who may have a range of comorbid conditions
Blood Glucose	Num	Concentration of glucose present in the blood (mmol/L)
Insurance	Num	Medical insurance expense coverage (Rs)
Salary	Num	Annual salary of the patient
FT/month	Num	Average number of foreign travels taken by the patient per month, considering the last 2 years
Infect_Prob	Num	Probability of the patient to get infected by SARS-CoV-2

Regarding transformations, we did a binary replacement for categorical attributes when it was possible, that is to say, in *Married* and *Gender*. Additionally, we replaced the categorical values of *Pulmonary_score*; instead of $\{<100, <200, <300, <400\}$, we utilized $\{100, 200, 300, 400\}$. As many models need to transform data before using it, we leveraged the *One Hot Encoder* from scikit learn [19] to encode the categorical attributes. We used this transformation as there is evidence that suggests that it provides better results than other state-of-the-art transformations [20]. Additionally, we utilized a logarithmic transformation on the numerical attributes as suggested in [21].

One of the objectives of this research was to predict the probability of infection for Set 2. However, as Set 2 does not include ground-truth values, we split Set 1 into two subsets: 1) 75% as for the training set, and 2) 25% as for the test set. We used the test set to evaluate the performance of a prediction model trained only with 75% of the observations with ground-truth values.

4.2 Experimental pipeline and results

In this section, we elaborate on every procedure of our experimental pipeline. We start by describing the feature reduction process, including MDS. Then, we present the specifications of our regression model. Later on, we outline our clustering technique. Along with these, we present our findings.

Feature reduction

We face a challenging analysis when we need to highlight the most important properties in a large volume of information by comparing all details. Removing information a priori may cause biased results [14]. Therefore, considering all the details compels adequate statistical and visualization methods capable of exposing the main features while ignoring those with low relevance. For this reason, having transformed and split the data in Set 1, we performed PCA for feature reduction. We started with a number of components $n = 44$, that is, the number of attributes. Then, we selected the number of components that add up to 1 variance-ratio, using the whitening parameter; although some information will be removed by the whitening, this improves the predictive accuracy of the estimators of downstream processes [19]. According to this procedure, it was necessary to keep only 35 out of 44 initial attributes. Afterwards, the data was reduced by PCA to 35 components, without the whitening parameter.

MDS

MDS analysis was performed on the complete, transformed Set 1 in order to mathematically cluster and visualize the data in a two dimensional space as suggested in [14]. This process translates the complex data to a geometric representation by working on the similarity matrices of distinct metrics. In this study, we use Euclidean distance to construct the similarity matrix; the plots of the MDS results are presented in Fig. 1. Since the standard MDS analysis is based on the distance of similarity or dissimilarity of the points, we could cluster the data by direct visualization of the plot. However, according to [14], this may be subjective. Thus, clustering algorithms are used to obtain a more precise result.

Regression model

In order to estimate the probability of infection for Set 2, we experimented on the training set and validated the model on the test set. The best regression method was selected by the R^2 score of a set of models, namely linear regression, ridge regression, Bayesian ridge regression, Hubber regression, lasso regression, bagging regressor, random forest regressor, AdaBoost regressor, linear support vector machine (SVM) regression and SVM RBF regression. Random forest regressor proved to be the most suitable model for our training dataset. Then, its parameters were iterated to get the highest regression score; the random forest regressor was trained with a maximum depth of 32 and all the features were considered when looking for the most desirable split.

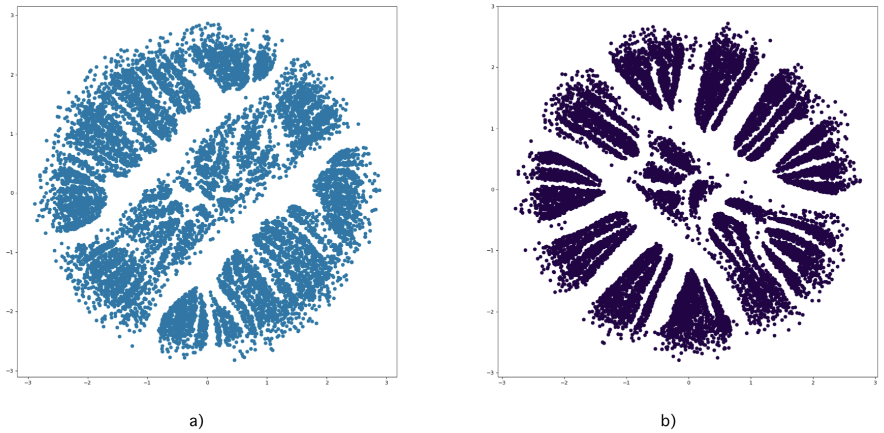

Fig. 1. Multidimensional Scaling on a) Set 1, and on b) Set 2.

Subsequently, we performed the same steps on the test set (categorical variables encoded, numerical variables transformed, reduced to 35 attributes by PCA). The R^2 score for this regression is of 0.2897. Up to this point, the R^2 score shows that the performance of the regression model is not competitive. According to [22], larger values of R^2 are an indication of a strong relationship between the independent and the dependent variable. Nevertheless, in practice, a small R^2 may indicate that the dependent variable might be partly influenced by other factors. Using the regression model generated by random forest regressor, the infection probability of Set 2 was predicted.

Clustering

The clustering process was conducted on the MDS results of Set 1 and Set 2. In order to have an overview of the distribution of the data in the MDS, we plotted the probability of infection of each observation associated with a color gradient (light tones are related to low probabilities and dark tones to high probabilities). Beforehand, the probability of infection was normalized; the results are presented in Fig. 2. These graphs set the basis to construct a clustering model to work on a intra-cluster probability analysis.

The algorithm of K-means was leveraged to perform the clustering analysis. K-means is a non-hierarchical clustering technique which groups elements into K clusters, defined by the user, so as to minimize the objective function given by the sum of the distances between the points and the centroids [14]. Nevertheless, good clustering is also subjective, hence, the clustering quality is measured with distinct indices. Here, we relied on the silhouette index. The silhouette values fluctuate in the interval [-1,1]; when the value is close to 1 or -1, the elements are assigned to the correct cluster. If the silhouette value is close to 0, then the elements could be assigned to another cluster [14]. Therefore, we plot the

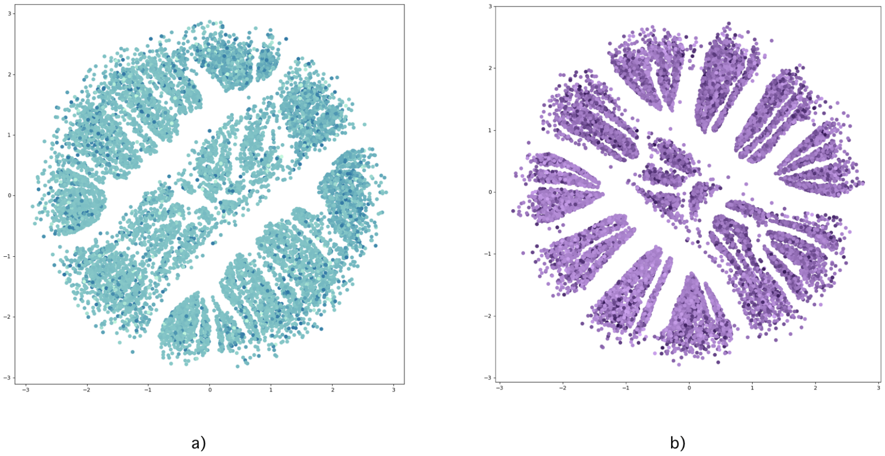

Fig. 2. Distribution of probability of infection in a) Set 1, and in b) Set 2. The probability values in Set 1 are ground-truth values from the dataset, whereas the values in Set 2 come from a regression model. Darker tones represent higher probabilites, and lighter tones otherwise.

silhouette average value against K number of clusters. As shown in Fig. 3, the best number of clusters K is 11 and 13 for Set 1 and Set 2, respectively. With the previous result we applied K-means to the MDS of Set 1 and Set 2.

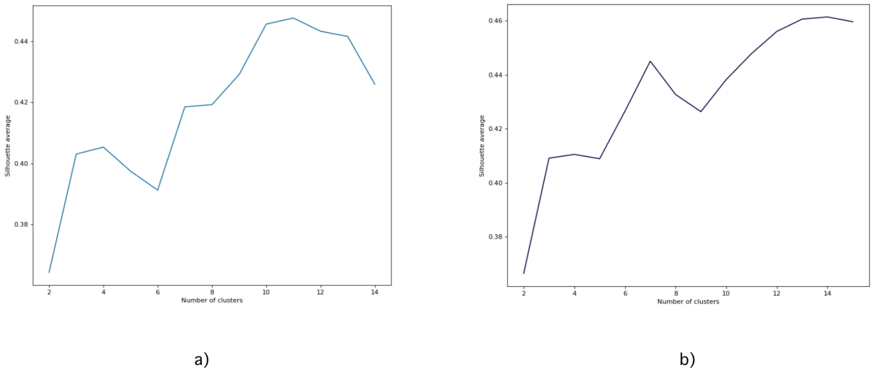

Fig. 3. K clusters vs. Silhouette average score for a) Set 1, and b) Set 2.

In Fig. 4 the resulting clusters are presented; the circles with the numbers mark the location of the centroids and each color represents a cluster. Moreover, the quality of the clustering was assessed by the silhouette analysis; as seen in Fig. 5, the plot of each cluster is above the mean and a low proportion of the points have the potential to be misclassified. Afterwards, we generated Fig. 6 for Set 1, and Fig. 6 for Set 2, where the infection probability of each observation by cluster is plotted; the color-probability relation is the same as stated above.

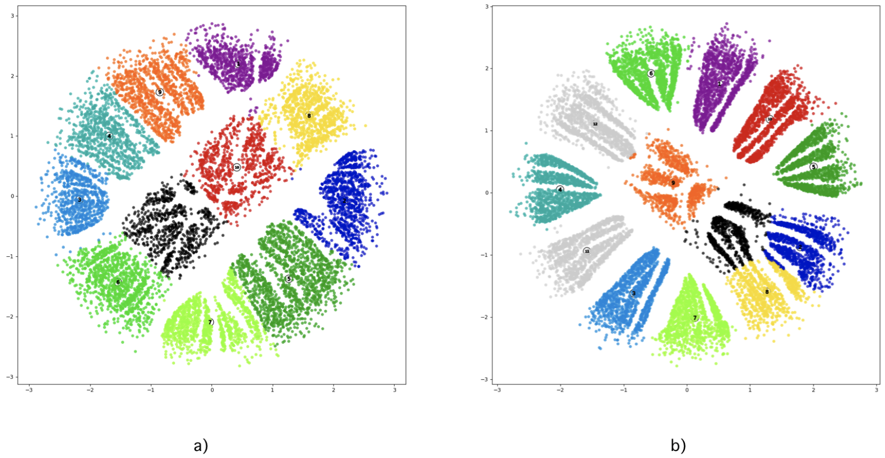

Fig. 4. Clusters obtained using K-means for a) Set 1, and b) Set 2.

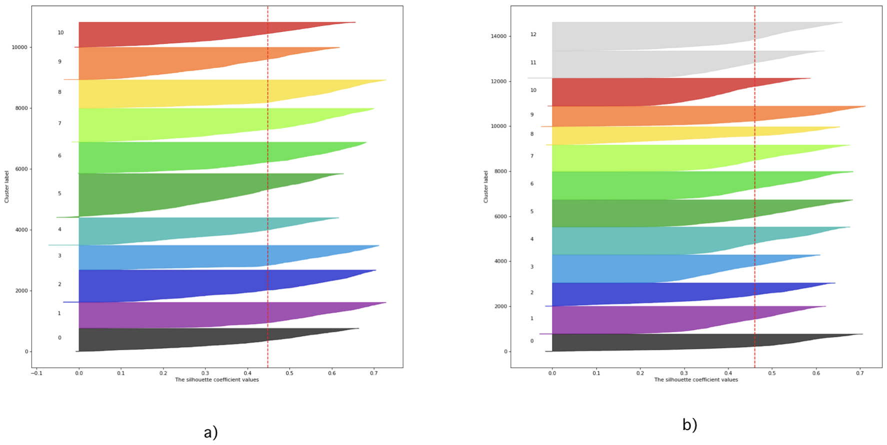

Fig. 5. Silhouette coefficient values vs. cluster for a) Set 1, and b) Set 2.

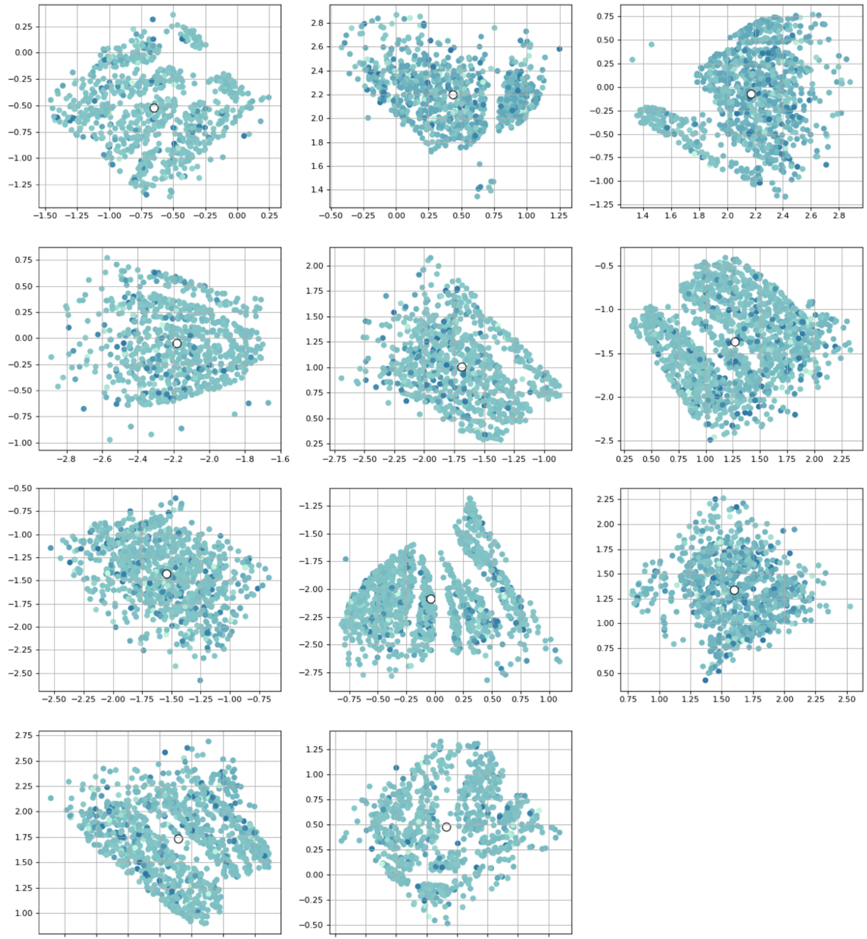

Fig. 6. Distribution of probability of infection per cluster for Set 1.

Statistical analysis

The attributes were compared across clusters with the one-way ANOVA tests and Welch's tests for those variables with heterogeneous variance. Furthermore, pairwise combinations were analyzed with t-tests adjusted for multiple comparisons with Tukey method and Benjamini and Hochberg method for variables with heterogeneous variance. The statistical analysis was performed in R (v3.6.2). The one-way ANOVA tests, Welch's tests and t-tests did not reveal any significant difference between the clusters of the pairwise combination of attributes of Set 1. However, the analysis performed on Set 2 reveal a very high significant ($p<0.001$) difference on nine attributes, namely region, comorbidity, pulmonary score, car-

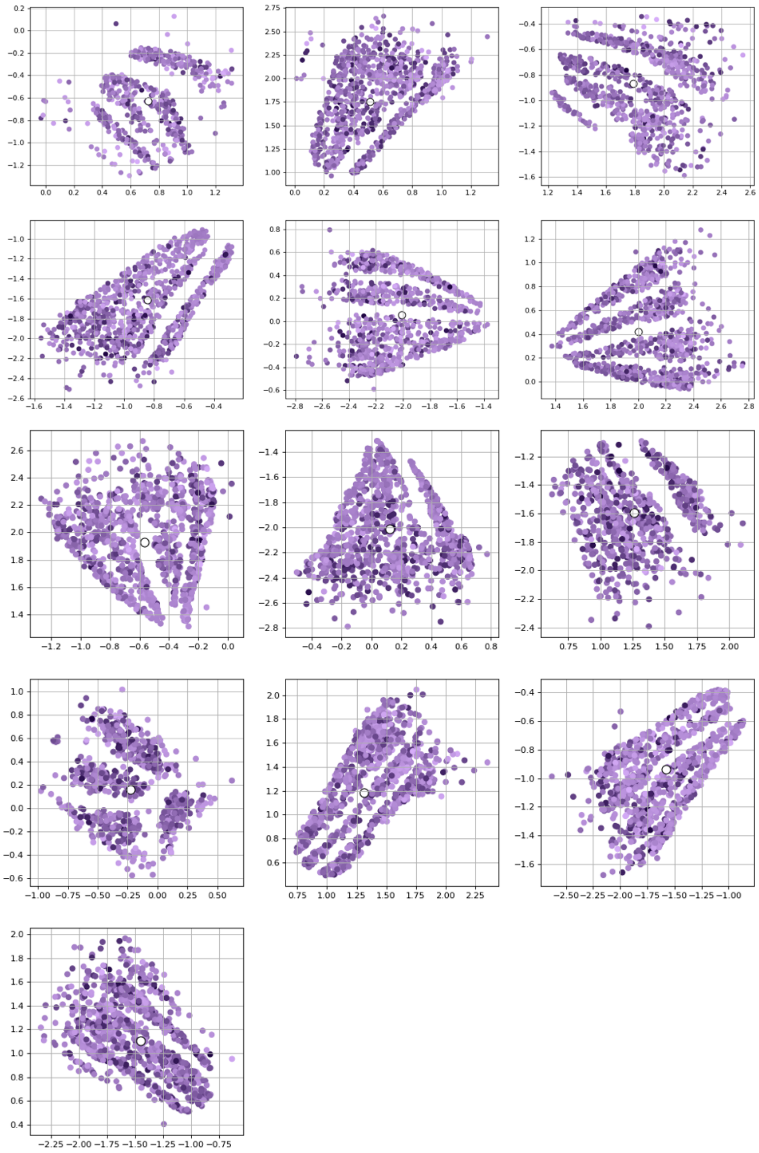

Fig. 7. Distribution of probability of infection per cluster for Set 2.

diological pressure, confirmed cases on the region per million inhabitants, deaths on the region per million inhabitants, Charlson index, insurance coverage and probability of infection. The most relevant attribute is pulmonary score, cardiological pressure and probability of infection ($p<2x10^{-16}$). According to the statistical results, cluster 1, 9 and 10 differ the most from the other clusters.

The mentioned attributes suggest that the probability of infection of SARS-Cov-2 is mostly related to pulmonary score. This concurs with other studies [3] [23] since SARS-CoV-2 mainly affects the respiratory system causing pneumonia and dypsnea (shortness of breath) [24]. Moreover, the model suggests that cardiological pressure is an important characteristic to take into consideration when evaluating the infection probability, which agrees with [25]. Also, it has been seen that older age, male gender and hyperglycemia increases the risk of infection [25]. However, our model did not find significant difference between clusters in these features. Comorbidity alongside the Charlson index (prediction of 10-year survival in patients with multiple comorbidities) does correlate with the probability of infection. Study [26] associated hypertension, diabetes, chronic obstructive pulmonary disease, cardiovascular disease and cerebrovascular disease as major risk factors for developing COVID-19. Nonetheless, the authors did not obtain correlation between COVID-19 and liver disease, malignancy or renal disease.

Our model was limited on the number of elements of the clinical test. Li et al. concluded that patients with high lactate dehydrogenase (LDH) level need early intervention to prevent a severe development of COVID-19 [25]. Furthermore, [7] proved by a interpretative model that LDH, lymphocytes, high-sensitivity C-reactive protien, eosinophils and monocytes levels can predict risk and mortality of the cases. On the other hand, it is important to emphasize that our model find that the region is an important factor when clustering the data. This suggests that COVID-19 may have distinct afflictions depending on the region, either by the diversity of inhabitants or by a molecular change in the virus. The mentioned proposition could be verified with further genomic analysis, more profound clinical tests and a broader area of study.

5 Conclusions

In this paper, we presented an experimental pipeline to extract information from a COVID-19 dataset. After preprocessing our data, we ended up with a 44-dimensional space. In this framework, our results show that MDS can be adopted as a visualization technique as it gives an insight into high-dimensional data and its spatial distribution. Furthermore, we leveraged clustering algorithms to extract information from the MDS representation. According to our findings, the clusters are actually related to the probability of infection. This raises many questions as to perform a deeper analysis on the attributes intra-cluster. Moreover, one-way ANOVA tests revealed inter-cluster information, that is, the attributes that vary most from one cluster to the other. These variations were found only on a subset of the data and they support some of the previous findings in other studies. All in all, this experimental pipeline allows to visualize the distribution of the data along with the attribute variations.

References

1. “Coronavirus cases.” [Online]. Available: https://www.worldometers.info/coronavirus/
2. K. Andersen, A. Rambaut, W. Lipkin, E. Holmes, and R. Garry, “The proximal origin of sars-cov-2,” *Nature Medicine*, vol. 26, pp. 450–452, 03 2020.
3. X. Yang, Y. Yu, J. Xu, H. Shu, J. Xia, H. Liu, Y. Wu, L. Zhang, Z. Yu, M. Fang, T. Yu, Y. Wang, S. Pan, X. Zou, S. Yuan, and Y. Shang, “Clinical course and outcomes of critically ill patients with sars-cov-2 pneumonia in wuhan, china: a single-centered, retrospective, observational study,” *The Lancet Respiratory Medicine*, vol. 8, pp. 475–481, 02 2020.
4. WHO, “Global surveillance for covid-19 caused by human infection with covid-19 virus: interim guidance, 20 march 2020,” World Health Organization, Technical documents, 2020.
5. N. Peiffer-Smadja, R. Maatoug, F.-X. Lescure, E. D’Ortenzio, J. Pineau, and J.-R. King, “Machine learning for covid-19 needs global collaboration and data-sharing,” *Nature Machine Intelligence*, 05 2020.
6. I. Apostolopoulos and M. Tzani, “Covid-19: Automatic detection from x-ray images utilizing transfer learning with convolutional neural networks,” *Physical and Engineering Sciences in Medicine*, 04 2020.
7. L. Yan, H.-T. Zhang, J. Goncalves, Y. Xiao, M. Wang, Y. Guo, C. Sun, X. Tang, L. Jing, M. Zhang, X. Huang, Y. Xiao, H. Cao, Y. Chen, T. Ren, F. Wang, Y. Xiao, S. Huang, X. Tan, N. Huang, B. Jiao, C. Cheng, Y. Zhang, A. Luo, L. Mombaerts, J. Jin, Z. Cao, S. Li, H. Xu, and Y. Yuan, “An interpretable mortality prediction model for covid-19 patients,” *Nature Machine Intelligence*, vol. 2, pp. 283–285, 2020.
8. M. Kubat, *Introduction to Machine Learning*. Springer, 2017.
9. R. M. Carrillo-Larco and M. Castillo-Cara, “Using country-level variables to classify countries according to the number of confirmed covid-19 cases: An unsupervised machine learning approach,” *Wellcome Open Research*, vol. 5, no. 56, p. 56, 2020.
10. H.-P. Kriegel, P. Kröger, J. Sander, and A. Zimek, “Density-based clustering,” *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 231–240, 2011.
11. F. Song, Z. Guo, and D. Mei, “Feature selection using principal component analysis,” in *2010 International Conference on System Science, Engineering Design and Manufacturing Informatization*, vol. 1, 2010, pp. 27–30.
12. S. Khalid, T. Khalil, and S. Nasreen, “A survey of feature selection and feature extraction techniques in machine learning,” in *2014 Science and Information Conference*, 2014, pp. 372–378.
13. I. Borg and P. J. Groenen, *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.
14. A. M. Lopes, J. P. Andrade, and J. T. Machado, “Multidimensional scaling analysis of virus diseases,” *Computer methods and programs in biomedicine*, vol. 131, pp. 97–110, 2016.
15. S. Dolgikh, “Identifying explosive cases with unsupervised machine learning,” *medRxiv*, 2020.
16. R. Garg, “Coronavirus disease 2019 (covid-19) india,” 03 2020. [Online]. Available: https://www.kaggle.com/rahulgarg28/coronavirus-disease-2019-covid19-india

17. P. Royston, “Multiple imputation of missing values,” *The Stata Journal*, vol. 4, no. 3, pp. 227–241, 2004.
18. S. van Buuren and K. Groothuis-Oudshoorn, “mice: Multivariate imputation by chained equations in r,” *Journal of Statistical Software, Articles*, vol. 45, no. 3, pp. 1–67, 2011.
19. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, “Scikit-learn: Machine learning in Python,” *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.
20. K. Potdar, T. Pardawala, and C. Pai, “A comparative study of categorical variable encoding techniques for neural network classifiers,” *International Journal of Computer Applications*, vol. 175, pp. 7–9, 10 2017.
21. H. Wickham, “Tidy data,” *The American Statistician*, vol. 14, 09 2014.
22. A. Siegel, “Chapter 11-correlation and regression: Measuring and predicting relationships, practical business statistics,” 2016.
23. F. M. Beloncle, B. Pavlovsky, C. Desprez, N. Fage, P.-Y. Olivier, P. Asfar, J.-C. Richard, and A. Mercat, “Recruitability and effect of peep in sars-cov-2-associated acute respiratory distress syndrome,” *Annals of intensive care*, vol. 10, pp. 1–9, 2020.
24. A. Jain and D. J. Doyle, “Stages or phenotypes? a critical look at covid-19 pathophysiology,” *Intensive Care Medicine*, p. 1, 2020.
25. X. Li, S. Xu, M. Yu, K. Wang, Y. Tao, Y. Zhou, J. Shi, M. Zhou, B. Wu, Z. Yang *et al.*, “Risk factors for severity and mortality in adult covid-19 inpatients in wuhan,” *Journal of Allergy and Clinical Immunology*, 2020.
26. B. Wang, R. Li, Z. Lu, and Y. Huang, “Does comorbidity increase the risk of patients with covid-19: evidence from meta-analysis,” *Aging (Albany NY)*, vol. 12, no. 7, p. 6049, 2020.